Iris classification

Aim:

The Iris Classification project involves creating a machine learning model to classify iris flowers into three species (Setosa, Versicolour, and Virginica) based on the length and width of their petals and sepals. This is a classic problem in machine learning and is often used as an introductory example for classification algorithms.

Problem Statement:

- The model should achieve a high level of accuracy in classifying iris species.
- The model's predictions should be consistent and reliable, as measured by cross-validation.
- The final report should provide clear and comprehensive documentation of the project, including all code, visualizations, and findings.

By achieving these objectives, the project will demonstrate the ability to apply machine learning techniques to a classic classification problem, providing insights into the characteristics of different iris species and the effectiveness of various algorithms for this task.

Steps:

1. Dataset Preparation

Few rows of dataset:

SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa

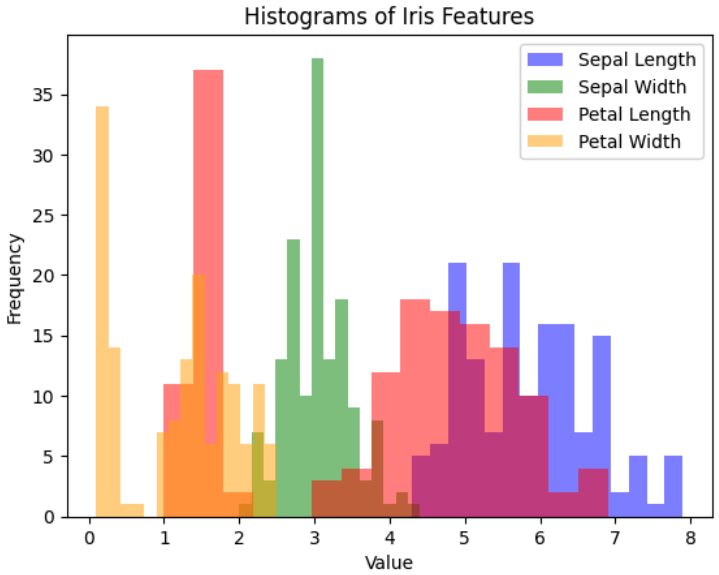
2. Data Exploration and Visualization

● Load the dataset and explore it using descriptive statistics and visualization techniques.

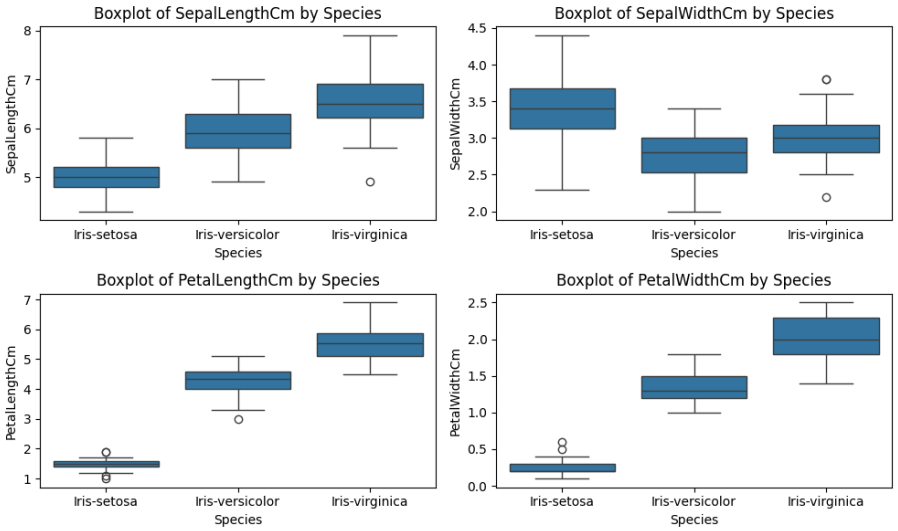
● Use libraries like Pandas for data manipulation and Matplotlib/Seaborn for visualization.

● Example visualizations include scatter plots, pair plots, and histograms to understand the distribution and relationships between features.

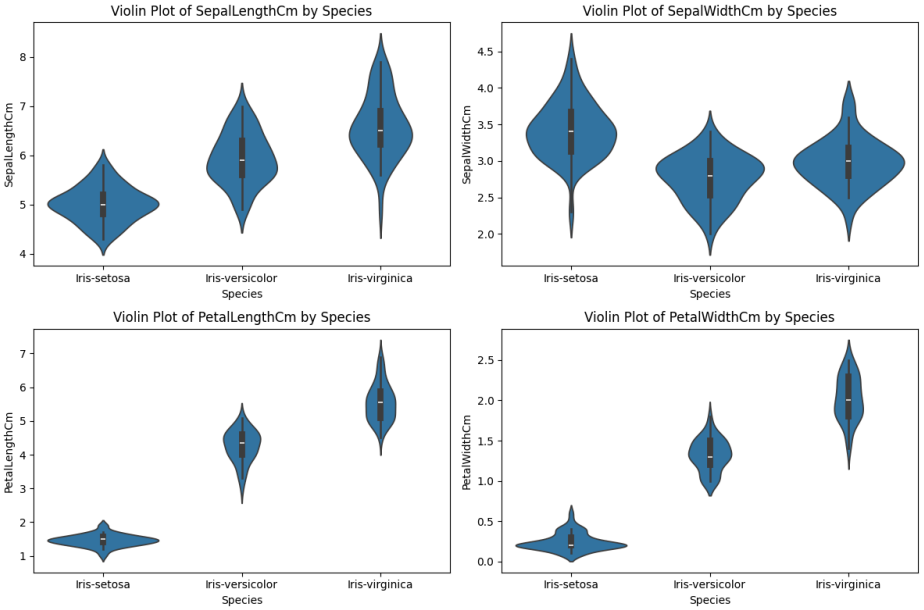
Following insights are found:



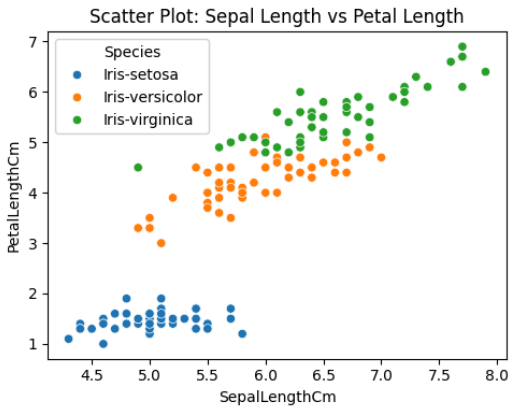
- Petal features (length & width): Strongest discriminators between species → clean separation.
- Sepal features (length & width): Show overlap, useful only for partial separation (Setosa vs others).
- The histograms confirm why petal measurements are preferred in ML models for Iris classification.

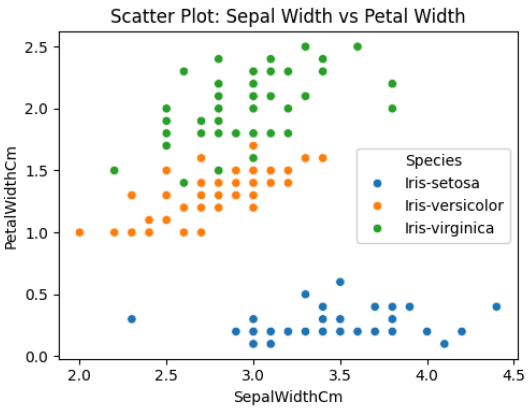


- Best features: Petal Length & Petal Width → clear separation of all species.
- Moderate feature: Sepal Length → helps but overlaps exist.
- Weakest feature: Sepal Width → not reliable due to strong overlap.
- Outliers: Present in Sepal Width (Setosa & Virginica) and Petal Width (Setosa).

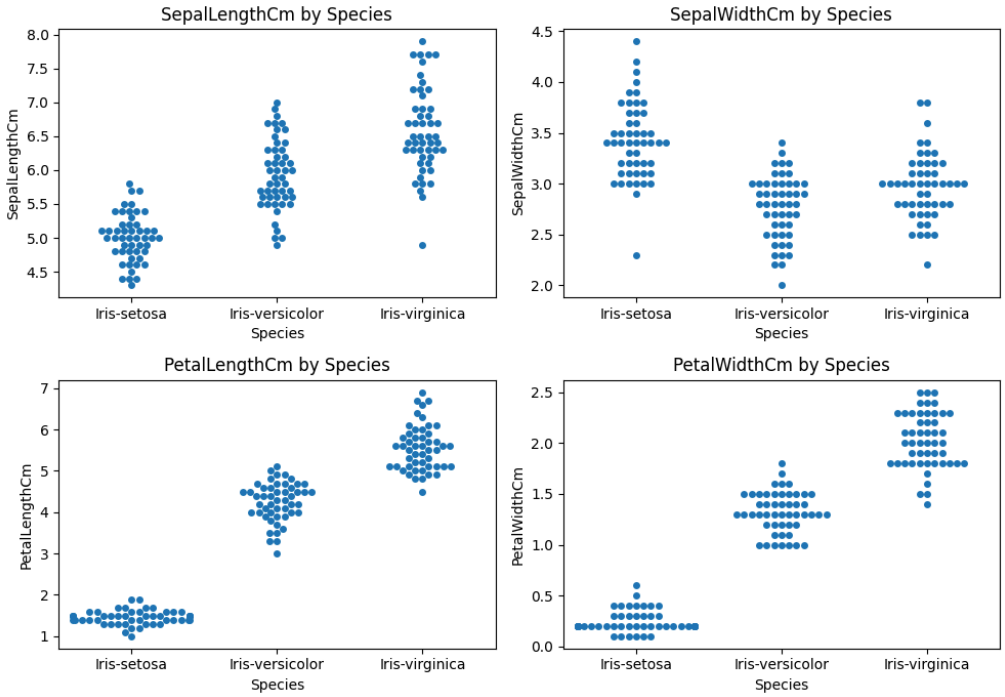


- Petal features (Length & Width) → Best predictors, clean separation.
- Sepal Length → Moderate use, overlaps for Versicolor & Virginica.
- Sepal Width → Least useful, high overlap across species.

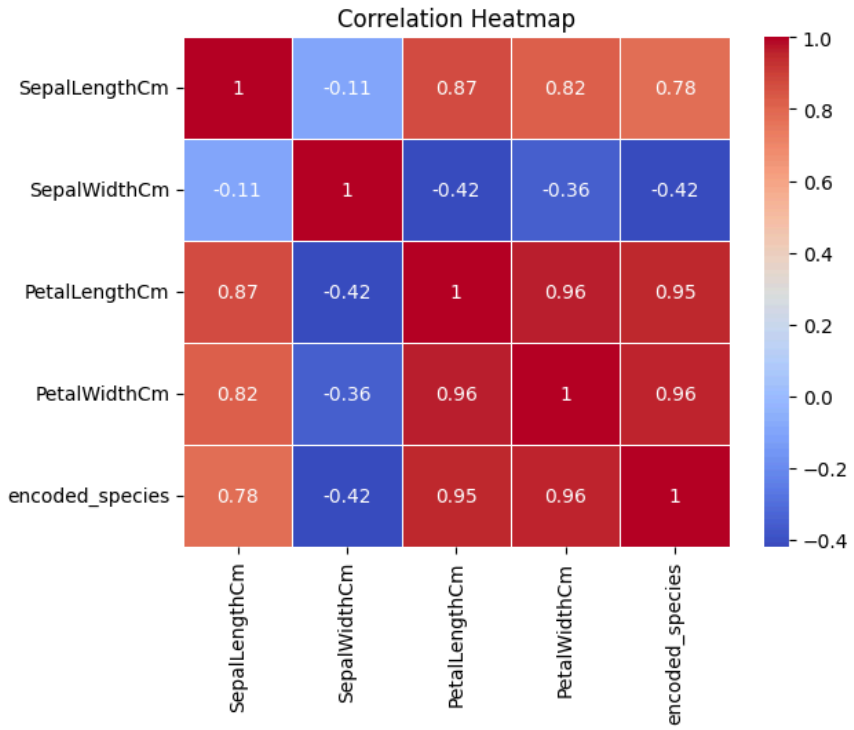




- Petal Length & Petal Width are the strongest distinguishing features across all species.
- Setosa is always clearly separated from Versicolor & Virginica.
- Versicolor vs Virginica: Overlap exists, making them harder to classify → but still separable using Petal features.
- Sepal features alone are weaker discriminators compared to Petal features.



- Best features: Petal Length and Petal Width (clear separation of species).
- Moderate feature: Sepal Length (helps distinguish Setosa, but Versicolor & Virginica overlap).
- Weakest feature: Sepal Width (heavily overlapping across species).
- Setosa: Always distinctly separated → easiest to classify.
- Versicolor & Virginica: Overlap slightly, explaining misclassifications in ML models.



3. Data Preprocessing

- Handled missing values.
- Standardize the features to ensure they are on a similar scale.
- Split the dataset into training and testing sets (commonly 80% training and 20% testing).

4. Model Selection and Training

Models used :

1. Support Vector Machine (SVM)
2. Logistic Regression
3. K-Means
4. PCA

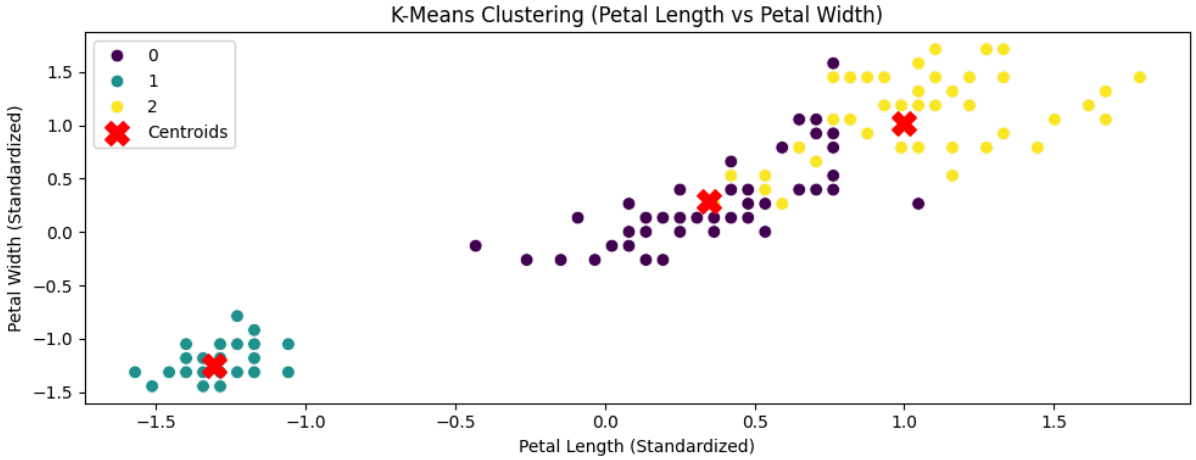
5. Model Evaluation

Model Comparison on Iris Dataset

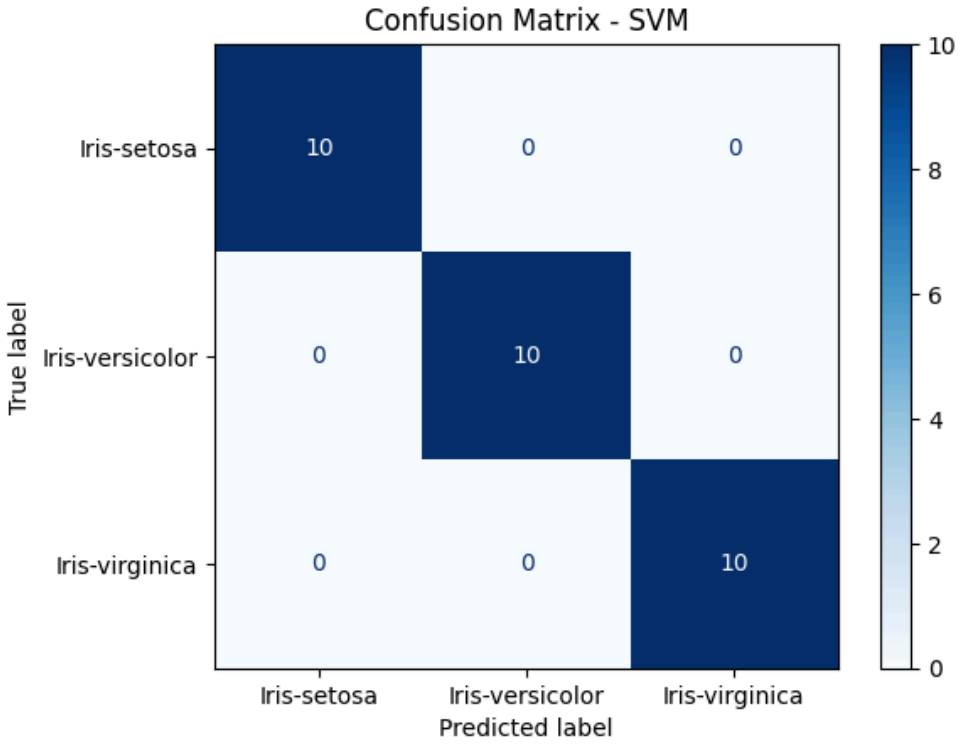
Model	Accuracy	Precision	Recall	F1-score	Notes	Explained Variance (PC1)	Explained Variance (PC2)	Total (PC1+PC2)
SVM	1.0	1.0	1.0	1.0	Supervised classifier	NaN	NaN	NaN

Logistic Regression	0.933333	0.933333	0.933333	0.933333	Supervised classifier	NaN	NaN	NaN
KMeans	NaN	NaN	NaN	NaN	Unsupervised clustering	NaN	NaN	NaN
PCA	NaN	NaN	NaN	NaN	Dimensionality reduction	0.727705	0.230305	0.95801

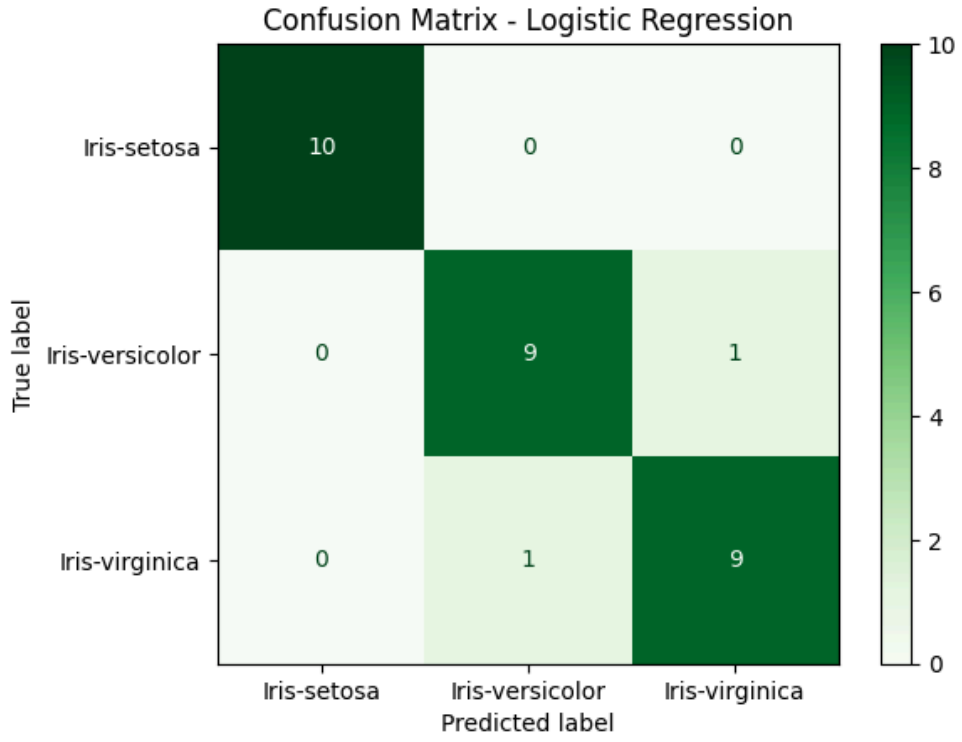
K-means:



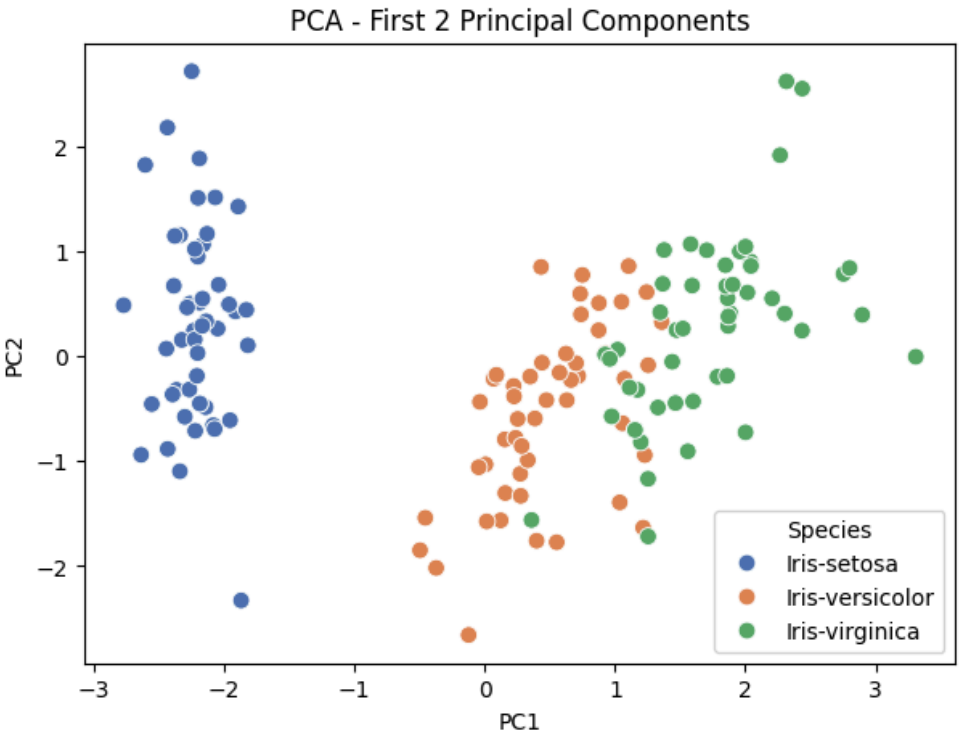
SVM



Logistic Regression:



PCA



6.Hyperparameter Tuning

Results:

Best Parameters: {'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}

Best Cross-Validation Accuracy: 0.975

Test Accuracy with Best SVM: 0.9333333333333333

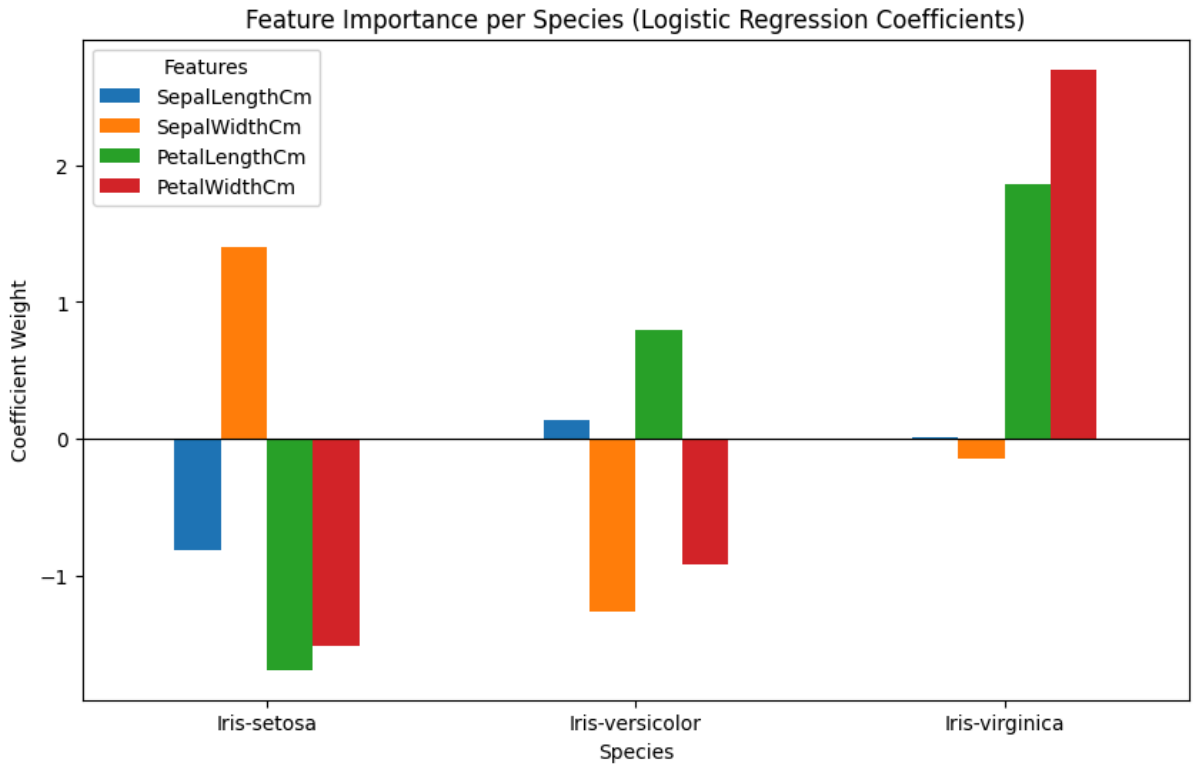
7. Model Interpretation and Insights

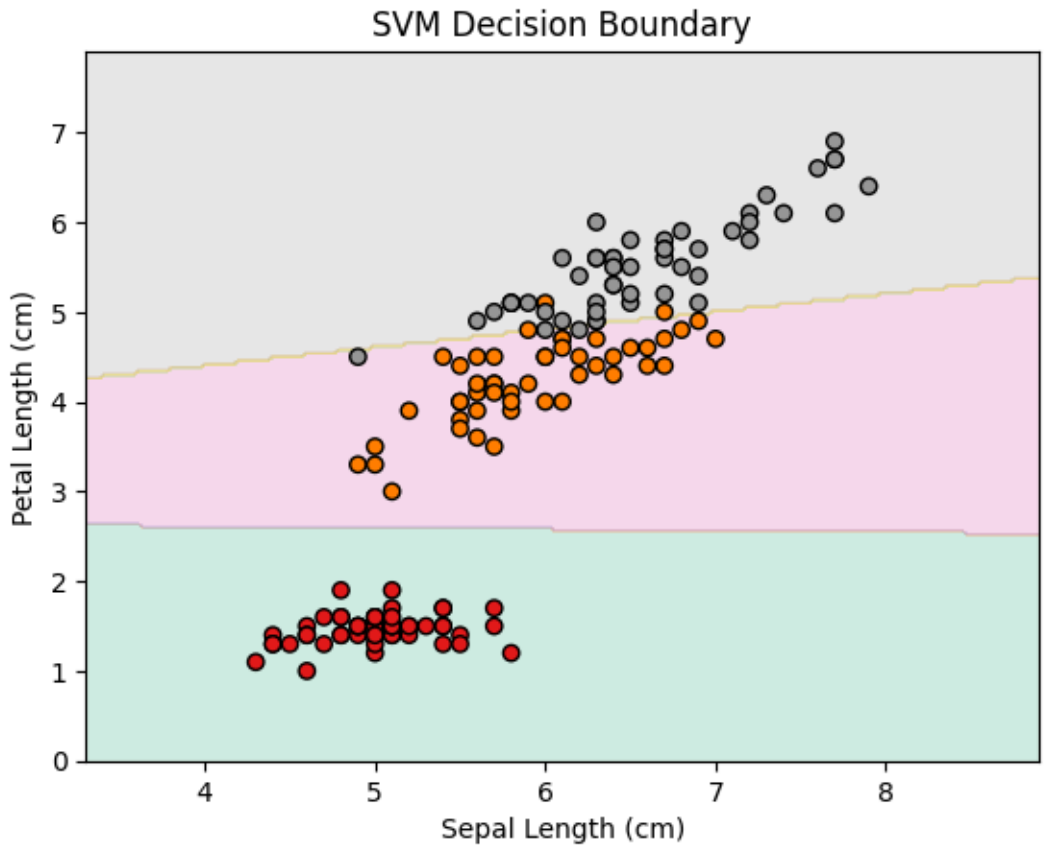
● Interpret the model results and understand which features are most important for the classification.

● Visualize decision boundaries if using models like Decision Trees or SVM.

Standardized Feature Means by Species

Species	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
Iris-setosa	-0.810166	1.393699	-1.687386	-1.518991
Iris-versicolor	0.130380	-1.246338	0.789195	-0.889440
Iris-virginica	0.012990	-0.144535	1.863173	2.698873





8. Conclusion

Petal width and petal length are the strongest predictors of iris species. Setosa is always well-separated, while Versicolor and Virginica are more challenging due to partial overlap.